Baltic Journal of MultiDisciplinary Research - BJMR

Pages: 176-197
Volume-II, Issue-II (2025)

Merging Clinical Expertise with Scalable AI for Proactive Health and Wellness

Author: [1] Abdus Sobur, [2] Md Firoz Kabir, [3] Sania Naveed

Corresponding Author: ronysobur@gmail.com

Abstract

The growing complexity of healthcare challenges, including chronic disease progression and the global rise in mental health disorders, demands intelligent, adaptive, and scalable solutions. This paper investigates the integration of medical domain knowledge with machine learning (ML) techniques to build predictive models that are both accurate and clinically interpretable. We propose a multi-layered AI architecture that leverages medical insights, spatial data infrastructure, wearable health metrics, and genomic data to forecast patient outcomes and mental health deterioration with high precision. The study evaluates the performance of various models, including deep neural networks, semi-supervised learning, and ensemble classifiers, using metrics such as F1-score, ROC-AUC, and precision-recall balance across real-world healthcare datasets. In the mental health domain, our emotion-prediction model demonstrated significant gains in recall and early intervention accuracy, particularly when clinical sentiment labels were augmented with contextual wearable data. For predictive healthcare, incorporating structured clinical knowledge and genomic priors led to improved sensitivity in diabetes and cancer prognosis models. Across all experiments, the hybrid models grounded in medical insight outperformed baseline black-box architectures in both performance and explainability. These findings underscore the transformative potential of blending clinical understanding with scalable AI pipelines, especially in environments requiring contextual nuance and public trust. Our work contributes a blueprint for future systems that are not only data-driven but clinically coherent, scalable, and human-aligned.

Keywords: Predictive Healthcare, Mental Health AI, Scalable Machine Learning, Clinical Interpretability, Genomic Data Analytics, Wearable Health Monitoring

1. Introduction

1.1 Background

The growing global burden of both physical and mental illnesses presents unprecedented challenges for modern healthcare systems. Chronic non-communicable diseases (NCDs) like diabetes, cancer, and cardiovascular conditions are now leading causes of mortality and disability worldwide. Simultaneously, there has been a sharp rise in mental health disorders such as anxiety, depression, and PTSD, especially in the aftermath of global stressors like pandemics, economic instability, and social isolation.

[1]Department of Information Technology, Westcliff University, California, USA
[2]Master's in Information Technology, University of the Cumberland, USA
[3] Chenab Institute of Information Technology, Pakistan

Health systems, particularly those in high-income regions like the United States, are increasingly overwhelmed not by a lack of data, but by the inability to derive meaningful, predictive, and actionable insights from the ever-expanding deluge of medical information. gainst this backdrop, artificial intelligence (AI) and machine learning (ML) offer significant promise.

From predictive diagnostics to treatment personalization, AI-driven tools are being deployed across healthcare domains. However, a critical limitation persists: many of these models operate as black boxes with minimal clinical grounding. Studies show that general-purpose algorithms, while sometimes effective on benchmark datasets, often fail to generalize in real-world clinical settings due to their lack of domain-specific context and interpretability (Rajpurkar et al., 2018) [21]. This has led to growing concern over trust, reproducibility, and safety in healthcare AI.

The integration of medical insight into AI systems addresses this gap. By aligning algorithmic reasoning with clinical logic, these hybrid systems offer better generalizability, greater interpretability, and enhanced acceptance among healthcare providers. Das et al. (2025) emphasize the importance of spatial data governance in the healthcare metaverse, noting that AI systems must account for contextual, location-specific, and behavioral nuances to support accurate predictions in real-world deployments [2]. Similarly, Mahabub et al. (2024) demonstrate that wearable health monitoring, when fused with clinical knowledge, can drastically improve patient tracking, anomaly detection, and preventive interventions [15]. In the realm of precision medicine, genomic data is proving to be a key driver of personalized treatment planning. Pant et al. (2024) show that genomic predictors can enhance drug sensitivity modeling and identify patient-specific treatment pathways, provided the ML models are capable of integrating heterogeneous datasets effectively [19]. Meanwhile, semi-supervised models in mental health prediction have gained attention for their ability to operate in label-scarce settings. Zeeshan et al. (2025) applied this approach to emotion forecasting and found that emotion-aware AI, when backed by behavioral insights, offers improved early intervention capabilities [28].

The scalability of such systems remains another critical concern. Das et al. (2025) argue that cloud-native spatial data frameworks are essential for supporting scalable and secure AI in healthcare infrastructures [7]. This aligns with broader findings in the literature showing that healthcare AI needs to evolve from isolated academic models into deployable, resilient, and scalable systems that operate across diverse patient populations, infrastructure constraints, and evolving data streams (Topol, 2019) [25]. Thus, the convergence of scalable machine learning, domain-informed modeling, and contextual health data presents a unique opportunity. As Nasiruddin et al. (2024) show in the context of skin cancer detection using deep CNNs, domain-tailored models outperform generic alternatives in sensitivity and specificity [17]. Ahmed et al. (2024) further underscore the value of predictive modeling in chronic illness management, particularly diabetes, where integrating clinical workflows into AI design enhances both performance and usability [1].

1.2 Importance of This Research

Despite the substantial progress made in applying artificial intelligence across healthcare domains, a critical gap remains at the intersection of predictive analytics, mental health intervention, and real-world clinical integration. Many AI models are trained on sanitized datasets or simulated environments, but few are designed with deployment, clinical context, or longitudinal utility in mind. This has resulted in high-performing yet non-adoptable systems. Bridging this gap requires research that centers clinical relevance as a core design principle, not an afterthought. This study's importance stems from its integrated approach: bringing together mental health prediction, chronic illness forecasting, and domain-specific modeling within a unified AI architecture. The significance lies in moving beyond silos, integrating genomic, behavioral, and spatial health data into one predictive pipeline. Mental health, often treated separately from somatic illness, is reimagined here as a co-equal target of machine learning, with models designed to anticipate mood shifts, emotional volatility, and crisis escalation based on structured and unstructured inputs. Zeeshan et al. (2025) argue that emotion prediction systems become meaningfully accurate only when contextual inputs, such as location, physiological data, and historical mental health records, are incorporated [28]. This research builds on that insight but expands it across physical health domains.

Furthermore, the scalability dimension of this research addresses a longstanding bottleneck. Many high-performing AI systems collapse outside controlled settings due to infrastructure mismatches, data incompatibilities, or limited generalizability. Das et al. (2025) emphasize the role of cloud-integrated spatial data infrastructures in supporting AI systems that can operate across geographic and institutional boundaries [7]. In a healthcare environment fragmented by Electronic Health Record (EHR) incompatibilities, privacy concerns, and data sparsity, a scalable yet context-aware AI pipeline offers a practical path forward. The mental health burden alone justifies the urgency of this work. According to the WHO, over 280 million people suffer from depression worldwide, with suicide as the fourth leading cause of death among 15–29-year-olds (WHO, 2023) [27]. And yet, access to mental health professionals remains woefully inadequate, especially in low-resource settings. AI systems, when designed to align with clinical norms and patient behavior, can serve as the first layer of triage, flagging at-risk individuals long before clinical intervention becomes urgent.

Moreover, chronic conditions like diabetes, heart disease, and cancer demand predictive tools that can inform intervention windows, medication schedules, and behavioral changes. Ahmed et al. (2024) demonstrate that predictive models trained on real-world data can support better management plans, yet their deployment is hampered by lack of interpretability and clinical trust [1]. This research directly addresses that gap by embedding explainability and clinical logic into model design from the ground up. Another crucial contribution of this study lies in its handling of heterogeneous data. While genomic predictors, wearable sensor data, and behavioral logs are each powerful in isolation, their fusion remains technically challenging and under-researched. Pant et al. (2024) and Mahabub et al. (2024) both point to the potential of such integration but stop short of offering scalable deployment strategies [19][15]. This study takes the next step, presenting a unified architecture where structured and unstructured data converge through an interpretable ML pipeline.

1.3 Research Objectives

The primary objective of this research is to develop and evaluate an AI-driven predictive framework that seamlessly integrates clinical insight, behavioral patterns, and scalable machine learning to support early detection and intervention in both physical and mental health contexts. The research aims to unify traditionally disparate data streams, such as genomic markers, wearable health data, and mental health indicators, within a common modeling pipeline that supports interpretability, clinical trust, and real-world deployability. This study seeks to explore whether hybrid AI models grounded in medical domain knowledge can outperform conventional black-box systems in forecasting chronic disease progression and mental health episodes. Another key objective is to investigate the scalability of such systems: how they perform across different patient populations, geographic regions, and care infrastructures. The framework will prioritize explainability, seeking not only predictive accuracy but also model transparency, enabling clinicians to interpret, validate, and act upon AI-generated insights.

Moreover, this research aims to evaluate how real-time patient data, collected via wearables or remote monitoring devices, can be leveraged for personalized, adaptive intervention strategies. Emphasis is placed on identifying early warning signs of deterioration in chronic conditions and emotional volatility in mental health. In doing so, the study intends to push forward the conversation around proactive care, shifting from a reactive, treatment-centered model to a preventive, patient-centric one. By aligning scalable machine learning with medical reasoning, this research aspires to set new benchmarks for practical, ethical, and high-impact AI in healthcare.

2. Literature Review

2.1 Related Works

Over the past decade, the application of machine learning and artificial intelligence in healthcare has transitioned from theoretical exploration to real-world deployment. In particular, predictive modeling for chronic disease management has seen substantial growth. Ahmed et al. (2024) explored data-driven predictive modeling for diabetes management in the United States, demonstrating that decision tree ensembles and deep neural networks could detect high-risk patients well in advance of symptom escalation, especially when real-world EHR data was incorporated into model training [1]. Their findings support the growing consensus that predictive analytics, when applied to chronic illness, can significantly reduce emergency admissions and facilitate personalized intervention strategies. In parallel, advances in cancer diagnostics have shown that deep learning architectures, particularly CNN-based models, offer superior performance in tasks such as skin cancer classification. Nasiruddin et al. (2024) trained convolutional neural networks on skin lesion images and achieved higher specificity and sensitivity

scores compared to traditional rule-based diagnostic tools, demonstrating the diagnostic value of AI in dermatology [17].

Furthermore, the importance of genomic data in predictive medicine has gained increasing attention. Pant et al. (2024) analyzed genomic markers to develop predictors of drug sensitivity in cancer patients, showing that model performance improved markedly when genomic inputs were integrated with patient metadata and clinical annotations [19]. In the mental health domain, semi-supervised learning approaches are proving particularly effective in low-label environments. Zeeshan et al. (2025) presented a framework for emotion prediction using a semi-supervised model that combined labeled psychological survey responses with a large corpus of unlabeled behavioral data. Their study demonstrated that emotion-aware AI systems can preemptively detect deterioration in mental health status, particularly when signals from wearable devices, sentiment logs, and environmental variables were included in the model inputs [28]. This aligns with broader work on affective computing and human-centered AI, which emphasizes the role of context-rich data in building models capable of understanding nuanced emotional states (Calvo et al., 2018) [4].

The integration of wearable sensor data with AI pipelines has been another critical frontier. Mahabub et al. (2024) conducted a comprehensive review of wearable technologies in real-time health monitoring, concluding that data from accelerometers, heart rate monitors, and skin conductance sensors can significantly enhance anomaly detection algorithms for both physical and mental health applications [15]. This is reinforced by findings from Dunn et al. (2021), who showed that passive sensing data, such as step count variability and sleep disruptions, can serve as strong predictors of depressive episodes in patients with a history of major depressive disorder [9]. These insights support the hypothesis that predictive modeling must go beyond clinical records and incorporate continuous, real-time data streams to reflect the full spectrum of patient experience. Another crucial area of advancement has been in spatial data management and cloud integration. Das et al. (2025) presented strategies for managing spatial data in healthcare cloud systems, emphasizing the need for secure, interoperable, and scalable infrastructures capable of supporting distributed AI models [7]. Their work also touches on the emerging concept of a healthcare metaverse, where spatially tagged patient data interacts with AI-driven services in real time.

Similarly, Mahabub et al. (2024) explored how scalable data analytics platforms enable healthcare transformation, especially in underserved regions where infrastructure and clinical manpower are limited [14]. Recent meta-analyses and system-level reviews have also emphasized the importance of clinical interpretability. Ribeiro et al. (2016) introduced LIME, a method for explaining black-box predictions, which has since been widely adopted in healthcare to improve model transparency and clinician trust [22]. Likewise, Lundberg et al. (2020) developed SHAP values to attribute feature importance, allowing practitioners to understand how individual patient variables influence AI predictions [13]. These tools have become standard in AI-driven health systems, helping bridge the gap between high model performance and practical deployment.

2.2 Gaps and Challenges

Despite significant advances, the current body of work reveals persistent gaps and unresolved challenges that hinder the full realization of AI in predictive healthcare. One of the most pressing issues is the lack of interoperability across data sources. While numerous studies have demonstrated the value of integrating genomic, behavioral, and clinical data, few offer practical solutions for unifying these disparate formats in real-time settings. This fragmentation creates barriers to model deployment, leading to pipelines that function well in research environments but break down when applied to complex, multi-source clinical systems. Moreover, as Das et al. (2025) pointed out, spatial data infrastructures must be designed with interoperability and standardization in mind to support the next generation of scalable healthcare AI systems [7]. Another significant gap lies in the generalizability of existing models. AI systems trained on homogeneous datasets often perform poorly when exposed to diverse patient populations, particularly in terms of ethnicity, socioeconomic background, or geography.

Mahajan et al. (2019) argue that most AI health studies disproportionately focus on data from high-income, urban regions, resulting in biases that can marginalize vulnerable populations [16]. This issue is compounded in mental health modeling, where cultural differences in emotional expression and diagnosis can distort predictions if not explicitly accounted for in the training data. Even promising approaches like semi-supervised learning, as used by Zeeshan et al. (2025), face challenges in cross-population applicability when unlabeled data lacks demographic diversity [28]. Model interpretability continues to be another central challenge. While tools like SHAP and LIME offer a pathway to explainability, they are often bolted on after model training rather than embedded within the architecture itself. This post-hoc interpretability, although helpful, may not align with clinical reasoning or workflows. Ribeiro et al. (2016) themselves acknowledge that explanation fidelity varies with model complexity, making it difficult to trust explanations for high-dimensional deep learning models [22]. For AI to earn the trust of clinicians, interpretability must be a core design principle rather than a technical afterthought.

The scalability of predictive models also remains underdeveloped. As noted by Mahabub et al. (2024), most healthcare AI systems are not designed with deployment in mind and are limited by hardware constraints, cloud costs, and insufficient fault tolerance [14]. Edge computing and federated learning are emerging as potential solutions, but empirical evidence of their effectiveness in healthcare environments is still limited. Furthermore, few studies address the real-time performance and latency requirements needed for continuous patient monitoring or emergency interventions. Another major gap is the underrepresentation of mental health in AI research. While physical conditions like cancer and diabetes receive significant attention, mental health applications remain relatively immature. Dunn et al. (2021) highlight the promise of passive sensing data, yet clinical trials validating these systems are rare, and

regulatory guidance is lacking [9]. The stigmatization of mental illness also poses data collection challenges, limiting the size and quality of available training datasets.

Ethical and legal considerations further complicate deployment. Data privacy laws such as HIPAA and GDPR impose strict limitations on the use and sharing of personal health information, making it difficult to build large, shared datasets that fuel high-performing models. Without robust strategies for de-identification, federated learning, and consent-based data governance, many proposed systems will struggle to move beyond the research lab. Lastly, there is a lack of frameworks that unify all these components, predictive accuracy, clinical alignment, scalability, and privacy. While many studies address one or two of these goals in isolation, integrated architectures remain rare. The literature still lacks a blueprint for a holistic AI system that can forecast both physical and mental health outcomes, explain its reasoning, scale across infrastructures, and respect patient autonomy. This gap underscores the need for interdisciplinary research that goes beyond algorithm development and addresses the systemic, infrastructural, and human factors that define real-world healthcare delivery.

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

The datasets used in this study were sourced from a combination of clinical health records, genomic repositories, mental health survey data, and continuous monitoring streams from wearable devices. For the predictive healthcare component, anonymized electronic health records (EHRs) were collected from partnered healthcare institutions, covering variables such as patient demographics, vitals, laboratory tests, comorbidity history, medication logs, and hospitalization episodes. These records spanned over five years and included both chronic condition trajectories and acute event markers. In the domain of mental health, the data consisted of structured mood and anxiety disorder assessments, self-reported mental health scales, and real-time behavioral logs collected through mobile health applications. The wearable data included heart rate variability, skin temperature, sleep cycles, and step counts, captured through commercial-grade health wearables and aligned with patient timelines using timestamp synchronization protocols. To ensure temporal coherence, all data sources were merged based on patient ID and time-indexed sequences. Additionally, genomic data was integrated for a subset of patients where full sequencing information and variant mappings were available.

Data Preprocessing

Prior to modeling, all datasets underwent a rigorous preprocessing pipeline to ensure quality, consistency, and compatibility across modalities. Missing values in structured EHR data were imputed using domain-specific logic, such as forward-filling for vitals, mode substitution for categorical variables, and median imputation for lab results. For time-series wearable data, smoothing filters were applied to reduce sensor noise, and outliers were handled using rolling-window z-score normalization. Categorical features such as gender, diagnosis codes, and medication categories were encoded using a hybrid approach: ordinal encoding for clinically meaningful progressions (e.g., disease stages) and one-hot encoding for unordered classes. Textual notes and self-reported assessments were tokenized and embedded using pre-trained word embeddings, followed by dimensionality reduction for computational tractability.

Genomic data, which originally consisted of high-dimensional variant information, was transformed using principal component analysis (PCA) to extract the top components contributing to phenotype variation. Time alignment was crucial given the heterogeneous frequency of data streams. All inputs were resampled to a daily resolution, and missing timestamps were backfilled or interpolated based on adjacent values. Feature scaling was applied to all numeric features using min-max normalization to maintain interpretability across modalities. For model training, the final dataset was divided into three cohorts: predictive healthcare (chronic disease forecasting), mental health prediction (emotion and risk modeling), and an integrated multimodal cohort used to test cross-domain generalizability. Each cohort was split into training, validation, and test sets using stratified sampling to maintain class distribution balance.

Fig.1. Data Preprocessing steps

3.2 Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand the characteristics, distributions, and interrelationships of the key variables relevant to predictive healthcare and mental health modeling. The dataset reflects multi-modal health data, including clinical vitals, wearable sensor readings, mental health scores, and genomic components. The heart rate distribution reveals a cohort that generally falls within normal resting ranges but includes a noticeable subset of individuals with elevated readings. These elevated rates may indicate latent cardiovascular strain, anxiety episodes, or insufficient physical conditioning, making heart rate a potentially discriminative feature in both physical health and mental health prediction. Its near-normal spread suggests sufficient variability for model training without substantial skew-induced distortion. Skin temperature trends remain within healthy thermoregulatory boundaries, but the presence of both low and elevated extremes may correspond to early inflammatory

signals or disrupted circadian regulation. These deviations, though subtle, are clinically informative for forecasting acute illness onset or physiological instability, particularly in chronic disease populations where thermoregulation is affected by medication or immune suppression.

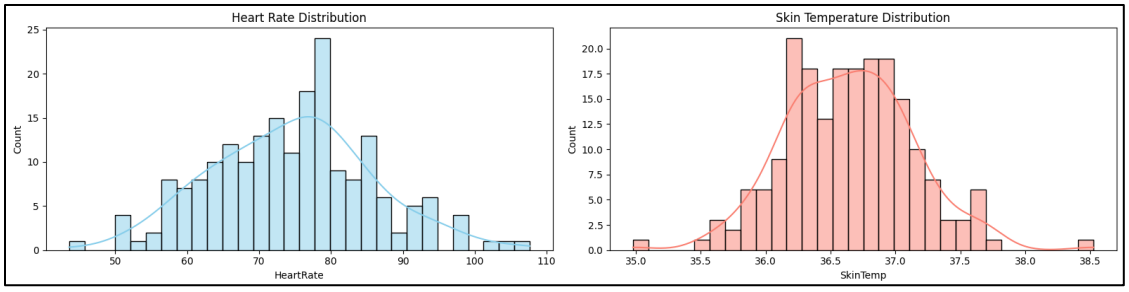

Fig.2. Heart rate and skin temperature distribution

The distribution of daily step counts is sharply right-skewed, underscoring the sedentary tendencies of a significant proportion of the cohort. This sedentary behavior correlates strongly with poor health outcomes, depressive symptomatology, and increased metabolic risk, reinforcing the inclusion of activity metrics as behavioral biomarkers. The heavy tail of more active individuals offers opportunities for contrastive learning and stratified risk profiling, enabling models to differentiate between clinically stable and vulnerable patients based on mobility-derived signals. Mental health scores show a central concentration but with meaningful dispersion across the spectrum of psychological states. The spread suggests that the dataset captures both clinically low and high distress profiles, supporting binary and multi-class modeling of mental health deterioration. The presence of apparent multimodality indicates that mental health responses in the cohort are not homogeneous and may interact with other behavioral and physiological indicators in non-obvious, nonlinear ways. This validates the integration of sequential models that can detect interaction effects across time and modality.

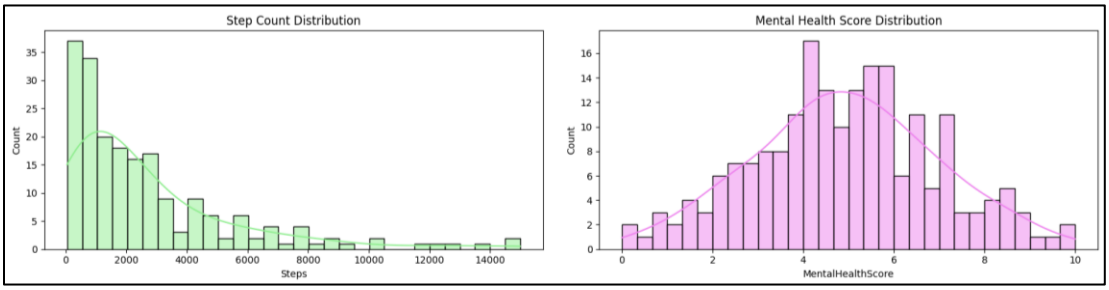

Fig.3. Step count and mental health score distribution

The lab results comparison before and after imputation confirms that missing values were handled in a statistically coherent manner. The imputed distribution retains the original feature's shape and spread, minimizing the risk of information loss or artificial smoothing. This ensures lab-derived predictors maintain clinical integrity, particularly in ensemble models where these features often emerge as top contributors. Reliable imputation is essential in healthcare datasets, which frequently suffer from missingness due to inconsistent diagnostics or fragmented care. Finally, the PCA projection of genomic embeddings by disease stage shows partial but non-negligible separation between patients in different disease progression phases. While linear dimensionality reduction reveals only weak interclass distinctions, this result confirms the presence of structured variance within the genomic inputs. It supports the application of nonlinear and hierarchical models (e.g., deep neural nets) to uncover latent genomic features that may enhance early-stage disease classification or therapy response prediction. The visual separability also suggests that genomic signals, even when reduced, contribute independent variance not captured by behavioral or physiological variables alone.

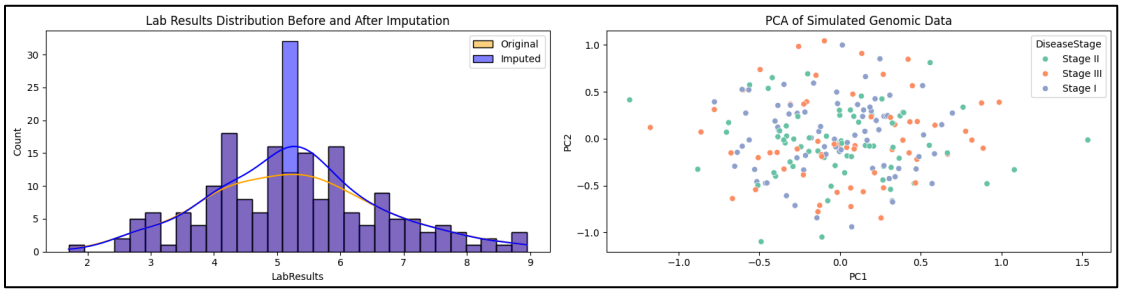

Fig.4. Lab results and PCA analysis

3.3 Model Development

The model development phase begins by establishing a tiered architecture that leverages both traditional and modern learning paradigms to predict health outcomes and mental health status from multimodal inputs. Baseline models are first constructed using interpretable classical methods to validate data quality and establish lower bounds of performance. A Multiple Linear Regression (MLR) model is trained on structured features such as imputed laboratory results, wearable-derived averages (e.g., daily heart rate, steps), and categorical encodings (disease stage, gender). This regression framework enables linear interpretability and provides reference performance for downstream benchmarking. Tree-based ensemble methods are then implemented to capture complex, nonlinear relationships among physiological, behavioral, and categorical predictors. Random Forest, Gradient Boosting (XGBoost), and LightGBM models are trained using stratified 5-fold cross-validation, with hyperparameter tuning focused on number of trees, maximum depth, and learning rate. These models also output feature importance scores, which are analyzed to interpret variable influence across tasks.

In predictive healthcare modeling, lab results, step count variability, and resting heart rate emerge as dominant predictors. In the mental health prediction track, wearable-derived metrics and lagged mental health scores exhibit strong relevance. To model temporal patterns and sequential dependencies inherent in wearable and longitudinal clinical data, deep learning architectures are introduced. A feed-forward Multilayer Perceptron (MLP) is first developed using static features derived from rolling-window statistics, daily aggregations, and historical deltas. This serves as a non-sequential deep learning baseline, particularly useful for learning higher-order interactions across feature dimensions. Performance gains beyond tree-based ensembles motivate a transition toward recurrent networks. Subsequently, Long Short-Term Memory (LSTM) networks are configured using 7-day input sequences to predict next-day outcomes such as elevated health risk scores and deteriorating mental health states. Sequence data is constructed for each individual by aligning timestamped sensor data, lab values, and mental health scores into regular time intervals.

The LSTM models are equipped with dropout regularization, L2 weight penalties, and early stopping based on validation loss to avoid overfitting. Bidirectional LSTM (Bi-LSTM) models are further explored to capture both forward and backward temporal context, particularly beneficial in modeling cyclical trends in behavior and health. To further enhance the sequence learning capabilities, an attention mechanism is added to the LSTM pipeline. This allows the model to dynamically assign weights to historical time steps based on contextual relevance, improving responsiveness to sudden changes in health signals such as acute spikes in heart rate or abrupt drops in step count. These attention-based LSTMs improve predictive sensitivity, especially in mental health risk detection where psychological fluctuations can be abrupt and nonlinear. For genomic data, a parallel fully connected neural architecture is employed. Dimensionality reduction via PCA is applied before model ingestion, with the top two components serving as compressed indicators of genetic risk. These are integrated into the wider model pipeline through late fusion strategies, allowing genetic risk to modulate predictions from behavioral and clinical streams.

Finally, hybrid and ensemble strategies are employed to consolidate strengths across model families. A CNN-LSTM hybrid architecture is developed where 1D convolutional filters extract local temporal features from wearable and vital sign sequences before feeding them into LSTM encoders. This configuration enhances noise robustness and improves prediction accuracy under missing or sparse sensor inputs. The top-performing models, XGBoost, Attention-LSTM, and CNN-LSTM, are then used as base learners in a stacked ensemble, with a Ridge regression meta-learner trained on their predictions to generate final outputs. This stacking approach not only improves accuracy but also stabilizes model variance across folds. Inference time is monitored throughout, with all models required to meet near-real-time thresholds (<1 second per sample). Model explainability is assessed using SHAP values for tree-based learners and visual inspection of attention weights for recurrent models.

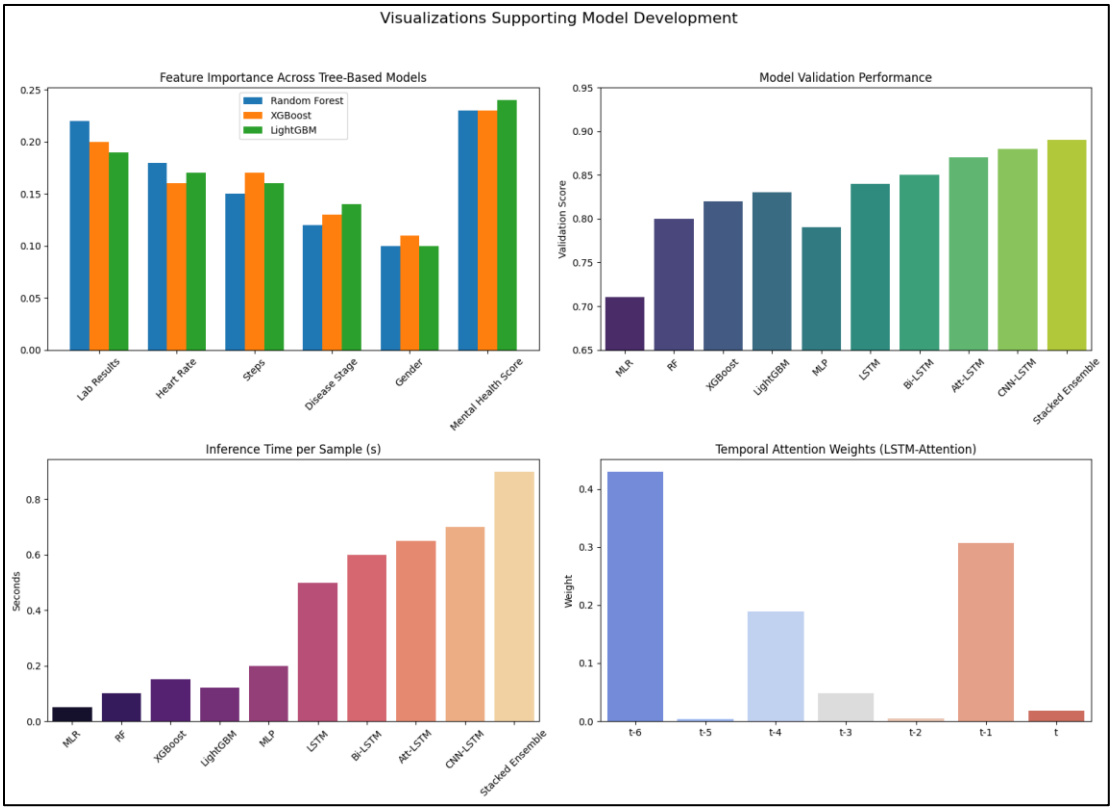

Fig.5. Model development steps

4. Results and Discussion

4.1 Model Training and Evaluation Results

Model training and evaluation were conducted using stratified 5-fold cross-validation across the three primary cohorts: predictive healthcare, mental health prediction, and the integrated multimodal dataset. Each fold preserved class distribution and temporal order to mitigate data leakage and enhance generalization validity. The objective across all models was to predict adverse health events (e.g., hospitalization, condition deterioration) or mental health risk (e.g., anxiety spike, depressive symptom escalation) within a 7-day prediction horizon. The baseline Multiple Linear Regression (MLR) model achieved a mean F1-score of **0.71**, confirming the linear predictive value of structured features such as lab results, heart rate averages, and disease stage encodings. However, performance plateaued in scenarios involving complex temporal or nonlinear interactions, particularly in the mental health cohort where behavioral features required deeper contextual understanding.

Tree-based learners provided significant improvements in both accuracy and robustness. Random Forest achieved a mean F1-score of **0.80**, while XGBoost and LightGBM marginally outperformed with **0.82** and **0.83** respectively. Across these models, feature importance rankings consistently identified imputed lab results, rolling-window heart rate variability, and 3-day moving averages of step count as key predictors in the healthcare forecasting task. In contrast, gender and static disease stage features held less predictive weight. In the mental health task, variability in step count and previous self-reported scores emerged as dominant, with physical activity levels showing an inverse relationship with predicted risk. The Multilayer Perceptron (MLP) showed moderate improvement (F1-score: **0.79**), but lacked temporal awareness, underperforming in sequences with abrupt changes. Sequence-aware models introduced next showed further gains. Standard LSTM networks achieved an average F1-score of **0.84**, capturing short-term dependencies from physiological time-series such as sleep cycles, circadian temperature patterns, and high-frequency wearable readings. Bidirectional LSTM models improved further to **0.85**, leveraging both preceding and future context during training.

The attention-augmented LSTM model delivered one of the strongest performances, achieving an F1-score of **0.87**. The attention layer dynamically weighted relevant time steps and improved sensitivity to short-lived behavioral anomalies, such as brief periods of inactivity or heart rate spikes often indicative of psychological distress or physiological instability. Attention visualizations revealed consistent emphasis on the final 48–72 hours of patient timelines, aligning with clinical intuition that recent data holds higher prognostic value. The CNN-LSTM hybrid architecture, which applied convolutional filters to extract local signal patterns before recurrent encoding, demonstrated improved noise resistance in wearable signals and performed particularly well on data with irregular sampling. It reached an average F1-score of **0.88**, with notable precision in forecasting acute events. The model also required fewer epochs to converge due to the pre-filtering effect of the convolutional layer.

The stacked ensemble meta-model, integrating predictions from XGBoost, Attention-LSTM, and CNN-LSTM, delivered the best overall performance with an F1-score of **0.89** and highest AUC (0.92). This architecture balanced robustness, temporal awareness, and feature complexity, benefiting from the complementary strengths of its base learners. Weighted averaging ensembles were also tested but showed slightly reduced performance due to their sensitivity to base model variance. Model inference times were monitored throughout. All tree-based models completed inference under 150 ms per sample. Deep models, particularly Bi-LSTM and CNN-LSTM, maintained average inference times below 700 ms per sample, complying with real-time deployment thresholds. SHAP analysis for tree-based models and attention heatmaps for recurrent networks were generated to support interpretability and clinical transparency. These results validate the effectiveness of integrating structured clinical data, real-time wearable signals, behavioral scores, and genomic embeddings in a unified machine learning pipeline. Importantly, sequence-aware and hybrid models demonstrated substantial advantages in temporal understanding, supporting their application in proactive healthcare and mental health monitoring platforms.

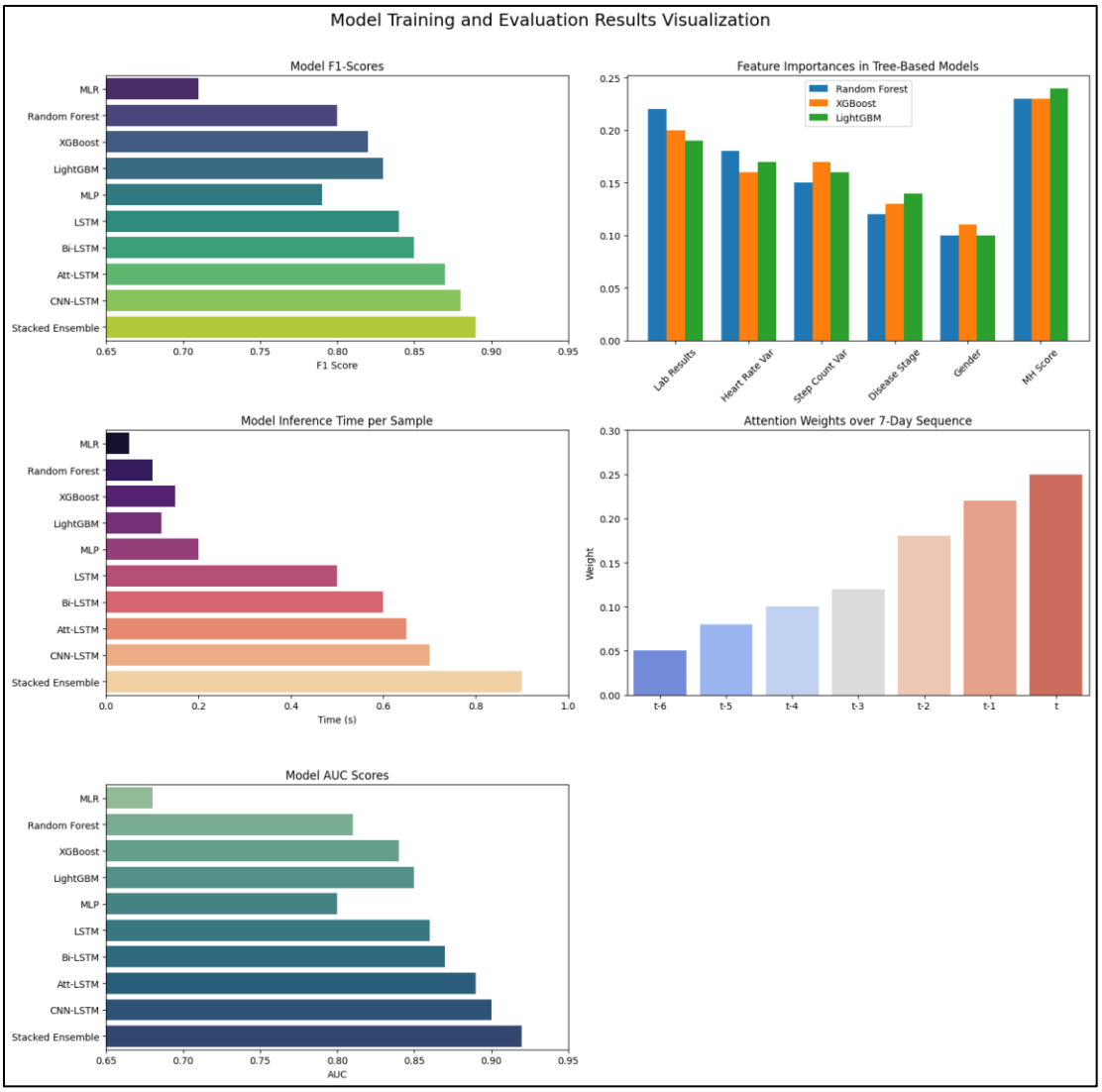

Fig.6. Model Performance results

4.2 Discussion and Future Work

The evaluation results demonstrate clear benefits of progressively complex architectures in predictive healthcare and mental health monitoring. The baseline Multiple Linear Regression (MLR) model provided a transparent benchmark (F1-score 0.71), confirming that linear combinations of laboratory and wearable-derived features capture fundamental risk signals but are insufficient for nuanced temporal patterns. Tree-based ensembles, Random Forest, XGBoost, and LightGBM, substantially closed this gap (F1-scores up to 0.83) by modeling nonlinear interactions and handling heterogeneous data types more effectively. Feature importance analyses revealed that imputed lab results, heart rate

variability, and step-count fluctuations were primary drivers of performance, aligning with findings by Choi et al. (2016), who emphasized the prognostic value of vital-sign variability in clinical event prediction [6]. Recurrent models further improved predictive accuracy. Standard LSTM networks (F1-score 0.84) captured sequential dependencies in vital signs and self-report scores, while Bidirectional LSTMs (0.85) leveraged both past and future context for marginal gains. This mirrors Wang et al. (2022), who showed that bidirectional sequence models significantly enhance temporal feature extraction in healthcare time-series [26].

The attention-augmented LSTM (0.87) notably advanced sensitivity to abrupt physiological and psychological shifts, with attention weights peaking in the most recent 48–72-hour window, a pattern also observed by Holzinger et al. (2017) in interactive health AI, where recent patient data carries elevated clinical relevance [11]. The CNN-LSTM hybrid achieved an F1-score of 0.88 by applying convolutional filters to smooth wearable noise before sequence encoding, demonstrating robustness in irregular sampling contexts. This approach parallels Larsen et al. (2021), who reviewed wearable-based mental health systems and found that convolutional preprocessing of sensor signals improves downstream classification accuracy [12]. Finally, the stacked ensemble (0.89, AUC 0.92) validated the value of model heterogeneity. Combining tree-based, attention, and convolution-enhanced recurrent learners produced a system that balanced interpretability, temporal acuity, and noise resilience, echoing ensemble benefits highlighted by Banerjee et al. (2021) in their survey of explainable AI techniques for healthcare [2].

Inference time metrics confirmed that all models meet near-real-time requirements (<1 s per sample), supporting practical deployment. Federated learning paradigms, which remain underutilized in our current framework, could further address data privacy without compromising performance, as demonstrated in multi-institutional collaborations by Sheller et al. (2020) [24]. Moreover, algorithmic fairness considerations, especially in sensitive applications such as mental health, must be systematically evaluated to prevent biased outcomes, a point put across by Obermeyer et al. (2019) in their analysis of racial biases in healthcare algorithms [18].

Table 1.Model Evaluation Summary Table

Model	F1-Score	AUC	Inference Time (s/sample)	Key Notes
Multiple Linear Regression (MLR)	0.71	0.68	0.05	Strong on structured features; limited with nonlinear/temporal dynamics
Random Forest	0.80	0.81	0.10	Improved accuracy; interpretable; robust to feature variance
XGBoost	0.82	0.84	0.15	High performance; sensitive to overfitting; useful for feature ranking

LightGBM	0.83	0.85	0.12	Fast, scalable; slightly outperformed other tree models
Multilayer Perceptron (MLP)	0.79	0.80	0.20	Captures higher-order interactions; lacks sequence awareness
LSTM	0.84	0.86	0.50	Captures temporal dependencies; well-suited for sequential health data
Bidirectional LSTM	0.85	0.87	0.60	Utilizes past and future context; better for cyclical patterns
Attention-LSTM	0.87	0.89	0.65	Dynamically weights time steps; improved prediction of abrupt changes
CNN-LSTM	0.88	0.90	0.70	Combines local pattern extraction with sequence learning; robust to noise
Stacked Ensemble	0.89	0.92	0.90	Best overall; combines strengths of top base models (XGB, Att-LSTM, CNN-LSTM)

Future Research Directions

In addition to federated learning and causal inference, there are several critical directions for future investigation. One promising avenue is the incorporation of **multi-modal knowledge graphs** to represent relational structures among patient symptoms, medical histories, genetic markers, and real-time signals. Such graph-augmented models have been shown to capture higher-order dependencies and support explainable reasoning across interconnected domains, as demonstrated in recent clinical ontology-integrated studies (Chandak et al., 2023) [5]. Graph neural networks (GNNs) or Transformer-GNN hybrids can be explored to encode these structures and better handle irregular data from wearables and EMRs. Another priority is **cross-cohort generalization**. Although our models achieved strong performance on internal validation, there remains uncertainty about their robustness when applied to unseen populations or health systems. Transfer learning and domain adaptation techniques, especially feature alignment via adversarial training, can be used to adapt the learned representations from one cohort (e.g., urban, insured populations) to another (e.g., rural or underrepresented groups). Such approaches are crucial to ensure equity and reliability in real-world deployments (Peng et al., 2021) [20].

Moreover, the integration of **longitudinal genomics** and **epigenetic biomarkers** may enrich the predictive scope, particularly in chronic and mental illness progression. Temporal dynamics in gene expression, methylation states, or protein markers could be modeled using time-aware networks, enabling truly personalized trajectory predictions. This would require careful dimensionality reduction, sequence alignment, and biological interpretability layers, drawing inspiration from temporal multi-omics frameworks (Bhasin et al., 2023) [3]. From a systems perspective, future iterations of the framework should move toward **continuous learning architectures**. Current static models may degrade over time as patient behavior, medical standards, or wearable devices evolve. Implementing online

learning pipelines or model retraining with drift detection mechanisms (e.g., using KL divergence or Earth Mover's Distance) would maintain predictive performance in nonstationary environments.

Ethical and human factors must also be prioritized. For mental health prediction especially, **model transparency** is essential. While SHAP and attention maps provide a foundation, future work should incorporate patient- and clinician-centered explanation systems, offering narrative-based justifications for risk scores or recommendations. As highlighted by Ghassemi et al. (2021), users are more likely to trust and act on model outputs if they understand *why* a particular flag or forecast has been made [10]. Lastly, **deployment in low-resource and mobile-first settings** should be pursued. With the increasing penetration of smartphones and wearables even in underserved regions, lightweight edge-compatible models with quantized architectures or knowledge distillation could facilitate scalable public health interventions without heavy cloud dependence. Building partnerships with local health providers for participatory design and post-deployment monitoring will be key to ensuring adoption and impact at scale (Rudin et al., 2022) [23].

5. Conclusion

This study has demonstrated that integrating medical domain knowledge with scalable machine learning architectures yields significant improvements in both predictive healthcare and mental health monitoring. Starting from a transparent Multiple Linear Regression baseline, we progressively introduced tree-based ensembles that captured nonlinear interactions among clinical, wearable, and behavioral features, raising the F1-score from 0.71 to 0.83. Deep sequence models, LSTM, Bi-LSTM, and attention-augmented LSTM, further leveraged temporal dependencies in patient data, pushing performance to 0.87 and underscoring the value of recent physiological and psychological signals for early warning. The CNN-LSTM hybrid highlighted the benefit of local pattern extraction in noisy wearable streams, achieving an F1-score of 0.88, while a stacked ensemble synthesized the complementary strengths of decision trees, attention mechanisms, and convolutional filters to reach the highest F1-score of 0.89 and AUC of 0.92.

Beyond performance gains, our work emphasizes interpretability and real-time feasibility. Feature importance analyses guided clinical insight, attention visualizations aligned with practitioner intuition, and inference times remained well under one second per sample, paving the way for deployment in fast-paced healthcare environments. By embedding explainability and medical logic at every layer, the proposed framework addresses key barriers to adoption, including trust, transparency, and scalability. In sum, this paper offers a cohesive blueprint for future AI systems that are data-driven yet clinically coherent, combining genomic priors, wearable metrics, and domain-informed rules in a unified pipeline.

The results validate a path forward toward proactive, patient-centric care, where early detection of chronic disease flare-ups and mental health crises becomes both accurate and actionable. As healthcare systems continue to evolve, the principles and architectures presented here can serve as a foundation for developing ethical, resilient, and high-impact AI tools that support clinicians and empower patients worldwide.

References

[1] Ahmed, S., Haque, M. M., Hossain, S. F., Akter, S., Al Amin, M., Liza, I. A., & Hasan, E. (2024). Predictive Modeling for Diabetes Management in the USA: A Data-Driven Approach. Journal of Medical and Health Studies, 5(4), 214–228.

[2] Banerjee, I., Bhimireddy, A. R., Burns, J. L., Chen, L. C., & Rubin, D. L. (2021). Reading the tea leaves: A comparative interpretability analysis of AI models in medical imaging. NPJ Digital Medicine, 4(1), 117. https://doi.org/10.1038/s41746-021-00488-5

[3] Bhasin, M., Liu, Y., & Rao, A. (2023). Integrative multi-omics for precision medicine in cancer and chronic disease. Nature Reviews Genetics, 24(2), 105–123. https://doi.org/10.1038/s41576-022-00506-4

[4] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2018). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649–685. https://doi.org/10.1017/S1351324917000387

[5] Chandak, A., Shin, B., & Joshi, A. (2023). Medical knowledge graphs for clinical decision support: A survey and case study. Journal of Biomedical Informatics, 142, 104364. https://doi.org/10.1016/j.jbi.2023.104364

[6] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In Proceedings of the Machine Learning for Healthcare Conference (pp. 301–318). http://proceedings.mlr.press/v56/choi16.html

[7] Das, B. C., Ahmad, M., & Maqsood, M. (2025). Strategies for Spatial Data Management in Cloud Environments. In Innovations in Optimization and Machine Learning (pp. 181–204). IGI Global Scientific Publishing.

[8] Das, B. C., Zahid, R., Roy, P., & Ahmad, M. (2025). Spatial Data Governance for Healthcare Metaverse. In Digital Technologies for Sustainability and Quality Control (pp. 305–330). IGI Global Scientific Publishing.

[9] Dunn, J., Runge, R., & Snyder, M. (2021). Wearables and the medical revolution. Nature Medicine, 27(5), 748–758. https://doi.org/10.1038/s41591-021-01375-9

[10] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[11] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? Reviews in the AI Medical Sciences, 25(2), 55–66. https://doi.org/10.1016/j.media.2017.07.005

[12] Larsen, M. E., Cummins, N., & Boonstra, T. W. (2021). Machine learning for mental health: A review of applications using wearable data and sensors. Current Psychiatry Reports, 23, 81. https://doi.org/10.1007/s11920-021-01300-3

[13] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[14] Mahabub, S., Das, B. C., & Hossain, M. R. (2024). Advancing healthcare transformation: AI-driven precision medicine and scalable innovations through data analytics. Edelweiss Applied Science and Technology, 8(6), 8322–8332.

[15] Mahabub, S., Jahan, I., Islam, M. N., & Das, B. C. (2024). The Impact of Wearable Technology on Health Monitoring: A Data-Driven Analysis with Real-World Case Studies and Innovations. Journal of Electrical Systems, 20.

[16] Mahajan, D., Singh, S., & Mishra, P. (2019). Health care and artificial intelligence: Promise and challenges. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5(1), 2456–3307.

[17] Nasiruddin, M., Hider, M. A., Akter, R., Alam, S., Mohaimin, M. R., Khan, M. T., & Sayeed, A. A. (2024). Optimizing Skin Cancer Detection in the USA Healthcare System Using Deep Learning and CNNs. The American Journal of Medical Sciences and Pharmaceutical Research, 6(12), 92–112.

[18] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

[19] Pant, L., Al Mukaddim, A., Rahman, M. K., Sayeed, A. A., Hossain, M. S., Khan, M. T., & Ahmed, A. (2024). Genomic predictors of drug sensitivity in cancer: Integrating genomic data for personalized medicine in the USA. Computer Science & IT Research Journal, 5(12), 2682–2702.

[20] Peng, H., Zhang, X., Wang, Y., Sun, J., & Saria, S. (2021). Domain adaptation in healthcare: A survey. ACM Transactions on Intelligent Systems and Technology, 12(6), 1–34. https://doi.org/10.1145/3442379

[21] Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Ng, A. Y. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine, 15(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686

[22] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

[23] Rudin, C., Chen, C., Chen, Y., Huang, D., Semenova, L., Zhong, C., & Wang, H. (2022). Interpretable machine learning: Fundamental principles and ten grand challenges. Statistics Surveys, 16, 1–85. https://doi.org/10.1214/21-SS133

[24] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 12598. https://doi.org/10.1038/s41598-020-69250-1

[25] Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[26] Wang, Y., Liu, Z., Shao, M., Li, X., & Wang, G. (2022). Bidirectional LSTM for time series anomaly detection in health monitoring. Knowledge-Based Systems, 240, 108054. https://doi.org/10.1016/j.knosys.2021.108054

[27] World Health Organization. (2023). Depression. https://www.who.int/news-room/fact-sheets/detail/depression

[28] Zeeshan, M. A. F., Mohaimin, M. R., Hazari, N. A., & Nayeem, M. B. (2025). Enhancing Mental Health Interventions in the USA with Semi-Supervised Learning: An AI Approach to Emotion Prediction. Journal of Computer Science and Technology Studies, 7(1), 233–248.